Substitution: A formal methods case study using monads and transformations

Françoise Bellegarde and James Hook*
Pacific Software Research Center
Oregon Graduate Institute of Science & Technology
PO Box 91000
Portland, Oregon 97291-1000
USA
{bellegar,hook}@cse.ogi.edu

April 7, 1994

Abstract

The specification and derivation of substitution for the de Bruijn representation of λ-terms is used to illustrate programming with a function-sequence monad. The resulting program is improved by interactive program transformation methods into an efficient implementation that uses primitive machine arithmetic. These transformations illustrate new techniques that assist the discovery of the arithmetic structure of the solution.

Introduction

Substitution is one of many problems in computer science that, once understood in one context, is understood in all contexts. Why, then, must a different substitution function be written for every abstract syntax implemented? This paper shows how to define substitution once and use the monadic structure of the definition to instantiate it on different abstract syntax structures. It also shows how to interactively derive an efficient implementation of substitution from this very abstract definition.

*The authors are supported in part by a grant from the NSF (CCR-9101721) and by a contract with Air Force Material Command (F19628-93-C-0069).

Formal methods that support reasoning about free algebras from first principles based on their inductive structure are theoretically attractive because they have simple and expressive theories. However, in practice they often lead to inefficient algorithms because they fail to exploit the "algebras" implemented in computer hardware. This paper examines this problem by giving a systematic program development and then describing a series of (potentially) automatic program transformations that may be used to achieve an efficient implementation.

The particular program development style employed is based on the categorical notion of a *monad*. This approach to program development has been advocated by Wadler[18, 19] and is strongly influenced by Moggi's work on semantics[15]. The substitution algorithm for λ-calculus terms represented with de Bruijn indexes serves as a case study.

The algorithm is mechanically transformed into first-order equations by previously published techniques. It is then refined to an equivalent first-order specification using first-order program transformation techniques. Finally the program is transformed to introduce standard arithmetic and boolean operators, thus achieving an efficient algorithm.

1 The Case Study: de Bruijn Representation

The de Bruijn representation of terms in the λ-calculus avoids the problems of bound variable names by using indexes to represent variables[5, 6]. The index assigned to an occurrence of a variable is the number of λ's in the abstract syntax tree between the occurrence and the λ that binds the variable. For example, the term:

$$\lambda u.(\lambda v.uv(\lambda w.uvw))(\lambda z.zu) \tag{1}$$

is represented by:

$$\lambda.(\lambda.1\,0(\lambda.2\,1\,0))(\lambda.0\,1) \tag{2}$$

This representation is most easily visualized by looking at the tree representing the term, which is given in Figure 1.

Free variables in λ-terms are also represented by indexes. The index of a free variable is an index greater than the number of λs on the path from the root to its occurrence. Two free variables are equal if, when placed in a context binding them, they would be equal as bound variables. For example, in (2), when the subterm $(\lambda.2\,1\,0)$ is considered out of context, indexes 2 and 1 represent distinct free variables.

The de Bruijn representation has the advantage that α-congruent λ-terms have identical representations. There is also no need to calculate sets of free and bound variables when performing substitution. Substitution is still not trivial, however, since indexes require adjustment as terms are moved into different binding contexts. This paper develops and refines a substitution algorithm for terms that use the de Bruijn representation.

To illustrate substitution, consider contracting the redex in (1). This yields

$$\lambda u.\, u(\lambda z.\, zu)(\lambda w.\, u(\lambda z.\, zu)w) \tag{3}$$

The contraction of the redex is expressed in the substitution calculus as:

$$[uv(\lambda w.\, uvw)](v \mapsto \lambda z.\, zu)$$

The substitution $(v \mapsto \lambda z.\, zu)$ is technically defined as a function from all variables to terms. It is more formally written as $(id \mid v \mapsto \lambda z.\, zu)$, that is, as the identity function perturbed at v to yield $\lambda z.\, zu$. This substitution will be called σ.

The contracted term (1) is represented by:

$$\lambda.\,0\,(\lambda.\,0\,1)\,(\lambda.\,1\,(\lambda.\,0\,2)\,0) \tag{4}$$

Figure 2 illustrates this term as a tree. Note that, after the contraction, the de Bruijn representation of the term that replaced v occurs with two distinct representations, $\lambda.\,0\,1$ and $\lambda.\,0\,2$, and

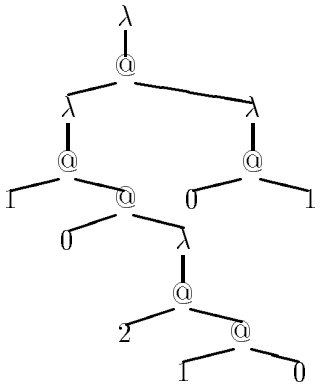

Figure 1: Tree representation of $\lambda u.(\lambda v.\, uv(\lambda w.\, uvw))(\lambda z.\, zu)$.

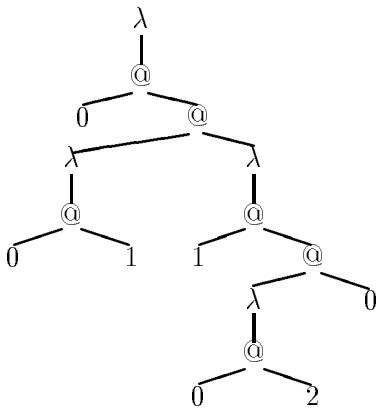

Figure 2: Representation of the contracted term.

the indexes associated with the occurrences of u within the scope of v in (2) are decremented in (4) because the λ binding v was removed in the contraction.

To express substitution directly with the de Bruijn representation these index adjustments must be defined. Earlier work of Hardin and Lévy and of Abadi and others on term calculi with explicit substitutions expressed these adjustments in terms of simple operations [1, 10, 9, 14]. Whenever a substitution enters a new binding context (i.e. a λ), the substitution function must be *shifted* to accommodate the new mapping of indexes to variables. For example, if 0 was mapped to $\lambda x.\, x$ (technically $\lambda 0$ in de Bruijn form) outside the λ, then inside the λ the index 1 must be mapped to $\lambda x.\, x$.

A second operation, *lift*, adjusts the indexes representing non-local variables, such as u in the example above. Every time a new binding context is entered, the value of the substitution function on every point in the domain must be lifted. Thus, the substitution σ expressed in de Bruijn notation is[1]:

$$\sigma(0) \quad = \quad \lambda 0\, 1$$

[1] The coercion of numbers to terms implicit here will become explicit in the programs developed below.

$$\sigma(n+1) \quad = \quad n$$

In the algorithm developed below an indexed family of functions is defined that gives the appropriately "lifted and shifted" substitution function for each binding context. This family will be generated inductively from the substitution σ. The first element of the family, σ_0, is the substitution σ. The second substitution in the family, σ_1, is obtained from σ_0 by shifting the domain and lifting all terms in the image. This gives $\sigma_1 0 = 0$, $\sigma_1 1 = \lambda.\, 0\, 2$ and $\sigma_1(n+2) = (n+1)$. In this case, these are the only substitutions needed, but in general any number may be required. The key to this development is to calculate this sequence of functions and then use a generic recursion scheme, such as that provided by the *map* function, that has been specialized to select the function from the family appropriate to the context.

The shifting transformation is easily captured by the approximate recurrence: $\sigma_{i+1} 0 = 0$ and $\sigma_{i+1}(n+1) \approx \sigma_i n$. To make it exact it is necessary to lift $\sigma_i n$. This is done by another sequence of functions:

$$\begin{aligned} f_0 n &= n+1 \\ \\ f_1 0 &= 0 \\ f_1(n+1) &= n+2 \\ \\ f_2 0 &= 0 \\ f_2 1 &= 1 \\ f_2(n+2) &= n+3 \end{aligned}$$

Observe that in the example a single application of f_1 to the body of $\sigma_1 1$ accounts for $\lambda.\, 0\, 1$ being lifted to $\lambda.\, 0\, 2$. In general the f_i are generated by $f_{i+1} 0 = 0$ and $f_{i+1}(n+1) = (f_i n) + 1$. Families of functions may be applied by a *map* functional that applies the ith member of the family to all indexes in the term in the scope of i λs (otherwise *map* leaves the structure of the term unchanged). Given *map*, the family of substitution functions, $(\sigma_0, \sigma_1, \ldots)$, is generated

by the initial substitution, σ_0, and the recurrence:

$$\begin{aligned} \sigma_{i+1}0 &= 0 \\ \sigma_{i+1}(n+1) &= map\,(f_0, f_1, \ldots)\,(\sigma_i n) \end{aligned}$$

The *map* function can be used again to apply the family of substitution functions, $(\sigma_0, \sigma_1, \ldots)$, to a term. This, however, results in terms of terms, since every variable has replaced its index by a term. This is not a problem, however, because the *Term* type constructor developed below is designed to be a monad; monads have a polymorphic function, *mult*, which performs the requisite flattening.

2 Monads

A *monad* is a concept from category theory that has been used to provide structure to semantics[15] and to functional programs[19]. In the computer science setting a monad is defined by a parametric data type constructor, T, and three polymorphic functions:

$$\begin{aligned} map &: (\alpha \to \beta) \to T\alpha \to T\beta \\ unit &: \alpha \to T\alpha \\ mult &: TT\alpha \to T\alpha \end{aligned}$$

The *map* function is required to satisfy:

$$\begin{aligned} map\ \ id_\alpha &= id_{T\alpha} \\ map\,(f \circ g) &= map\ f \circ map\ g \end{aligned}$$

The polymorphic functions *unit* and *mult* must satisfy:

$$\begin{aligned} mult_\alpha \circ unit_{T\alpha} &= id_{T\alpha} \\ mult_\alpha \circ (map\ unit_\alpha) &= id_{T\alpha} \\ mult_\alpha \circ mult_{T\alpha} &= mult_\alpha \circ (map\ mult_\alpha) \end{aligned}$$

A simple example of a monad is *list*. For lists, *map* is the familiar `mapcar` function of Lisp, *unit* is the function that produces a singleton list, and *mult* is the concatenate function that flattens a list of lists into a single list. Other examples of monads are given by Wadler[19].

Several categorical concepts are implicit above. The functional programming category has types as objects and appropriately typed functions between them as arrows. (Values are viewed as constant functions—arrows from the one element type. For the purpose of this paper types may be interpreted concretely as sets.) The requirements on *map* specify that the type constructor T and the *map* function together define a *functor*. The three laws given for *unit* and *mult* are the *monad laws*. By virtue of their parametric polymorphic types, *unit* and *mult* are *natural transformations*[17]. That is, they satisfy the following equational properties:

$$\begin{aligned} unit \circ f &= map\ f \circ unit \\ mult \circ map\ f &= (map(map\ f)) \circ mult \end{aligned}$$

Monads have been used to structure programs (and semantics) because it is often possible to characterize interesting facets of a specification as a monad. Algorithms to exploit the particular facet may frequently be expressed in terms of the *map*, *unit* and *mult* functions with no explicit details of the type constructors.

3 The Term Monad

3.1 Naive terms

Term structures and substitution are natural candidates for the application of monads. In this section monads are illustrated by terms without binding structure. In the next section the full substitution algorithm for de Bruijn terms is given in a monadic setting.

Consider the very simple term data type:

$$\begin{aligned} \textbf{datatype}\ Term'(\alpha) &= Var(\alpha) \\ &\mid App(Term'(\alpha) * Term'(\alpha)) \end{aligned}$$

It is easily verified that $Term'$ is a monad. The map, unit and multiplier, given in Figure 3, are easily calculated from the definition with the techniques of Hook, Kieburtz and Sheard[11].

$$
\begin{array}{lcl}
map\ f\ (Var\ x) & = & Var(f\ x) \\
map\ f\ (App(t,t')) & = & App(map\ f\ t, map\ f\ t') \\
unit & = & Var \\
mult\ (Var\ x) & = & x \\
mult\ (App(t,t')) & = & App(mult\ t, mult\ t')
\end{array}
$$

Figure 3: Definition of the map, unit and multiplier for $Term'$.

Taking the viewpoint that a substitution is a function from variables to terms, it is natural to associate the type $\alpha \rightarrow Term'(\beta)$ with a substitution function. It is then meaningful to apply the *map* function of the monad to a substitution, which yields a $Term'(Term'(\beta))$.

This intermediate "term-term" is the least intuitive aspect of the example. Essentially a term over type α has been converted to a term-term over β by replacing the α-values with β-terms, but not the *Var* constructors that had been applied to them. The multiplier function, which has the type $Term'(Term'(\alpha)) \rightarrow Term'(\alpha)$, is exactly what is needed to clean up this situation. In this case it removes the residual applications of the *Var* constructor in the term-term.

In summary, if σ is an appropriately typed substitution function, the action of the substitution on the simple term type above is given by:

$$mult \circ map\,\sigma$$

This use of the multiplier and the map together to obtain a function of type $T(\alpha) \rightarrow T(\beta)$ from a function f of type $\alpha \rightarrow T(\beta)$ is called the *natural extension* of f. Monads can be defined in terms of natural extension and unit.

3.2 Terms with binding

The development in Section 1 suggests that the specification of the substitution operation will be straightforward in a monadic data type with an appropriate *map*. The following type declaration extends the naive type above with λ-abstraction:

$$\begin{array}{lcl} \textbf{datatype}\ \mathit{Term}(\alpha) & = & \mathit{Var}(\alpha) \\ & | & \mathit{Abs}(\mathit{Term}(\alpha)) \\ & | & \mathit{App}(\mathit{Term}(\alpha) * \mathit{Term}(\alpha)) \end{array}$$

Note that even though the de Bruijn representation will use $\mathit{Term}(\mathit{Nat})$, *Term* is specified to be a parametric type constructor. This provides the structure to support the definition of *mult*, which will have type $\mathit{Term}(\mathit{Term}(\alpha)) \rightarrow \mathit{Term}(\alpha)$. It also gives *map*, *mult* and *unit* the polymorphic types that enable the assertion of the "free theorems" characterizing them as natural transformations[17].

As above, it is possible to automatically generate *map*, *mult* and *unit* functions for this type realizing a monadic structure. Unfortunately, the *map* function obtained with those techniques does not work with families of functions.

To accommodate the function sequences a new category, FunSeq, is used. The objects are data types, as before, but the morphisms are sequences of functions (formally $\mathbf{Hom}(A, B) = (B^A)^\omega$). Identities are constant sequences of identities from the underlying category; composition is pointwise, i.e. $(f_i)_{i\in\omega} \circ (g_i)_{i\in\omega} = (f_i \circ g_i)_{i\in\omega}$.

The *map* function for *Term* exploits the new structure by shifting the series of functions whenever it enters a new context. Its definition is given as a functional program:

$$\begin{array}{lcl} \mathit{map}\ (f_0, f_1, \ldots)\ (\mathit{Var}\ x) & = & \mathit{Var}(f_0\ x) \\ \mathit{map}\ (f_0, f_1, \ldots)\ (\mathit{Abs}\ t) & = & \mathit{Abs}(\mathit{map}\ (f_1, f_2, \ldots)\ t) \\ \mathit{map}\ (f_0, f_1, \ldots)\ (\mathit{App}(t, t')) & = & \mathit{App}(\mathit{map}\ (f_0, f_1, \ldots)\ t, \mathit{map}\ (f_0, f_1, \ldots)\ t') \end{array}$$

It is easily verified that $(\mathit{Term}, \mathit{map})$ satisfy the categorical definition of a functor.

Looking at these definitions, it is clear how to insert an ordinary function or value into the category, and it is straightforward to insert the families of functions needed for the example by

giving the initial element of the sequence and the functional that generates all others. Thus, one way to realize the *map* function of FunSeq in a functional programming setting is with the *map_with_policy* function introduced in Hook, Kieburtz and Sheard[11]:

$$\begin{array}{lcl} map\_with\_policy\ Z\ f\ (Var\ x) & = & Var(fx) \\ map\_with\_policy\ Z\ f\ (Abs\ t) & = & Abs(map\_with\_policy\ Z\ (Zf)\ t) \\ map\_with\_policy\ Z\ f\ (App(t,t')) & = & App(map\_with\_policy\ Z\ f\ t, \\ & & \quad map\_with\_policy\ Z\ f\ t') \end{array}$$

In this encoding Z is the functional that generates the sequence and f is the seed value. That is,

$$map\ (f, Zf, Z^2 f, \ldots) = map\_with\_policy\ Z\ f$$

The name *map_with_policy* refers to the notion of *policy function* introduced by Kieburtz[12, 11].

The *unit* and *mult* functions automatically generated for *Term* are:

$$\begin{array}{lcl} unit & = & Var \\ mult\ (Var\ x) & = & x \\ mult\ (Abs\ t) & = & Abs(mult\ t) \\ mult\ (App(t,t')) & = & App(mult\ t, mult\ t') \end{array}$$

These may be lifted to FunSeq by forming the trivial families $(unit, unit, \ldots)$ and $(mult, mult, \ldots)$. Simple induction proofs show that they satisfy the monad laws.

With these definitions in place the complete definition of substitution is given in Figure 4. Note that the algorithm makes no explicit mention of the data constructors. It only uses the information about the type implicit in the definition of *map_with_policy*, *unit* and *mult*. Appendix A proves this algorithm is correct with respect to the calculus of explicit substitutions of Curien, Hardin and Lévy[10, 9].

4 Transformation to a First-Order Set of Equations

To obtain a practical algorithm, the substitution function *apply_substitution* in Figure 4 must be made more efficient. This section shows how this transformation can be done automatically.

```
fun apply_substitution σ0 M =
    let  fun succ x = x + 1
         fun lift f
             = λn. if n = 0 then 0 else 1 + f(n − 1)
         fun shift σ
             = λn. if n = 0 then unit 0
                   else map_with_policy lift succ (σ(n − 1))
    in   mult(map_with_policy shift σ0 M)
    end
```

Figure 4: Substitution function

Program transformation systems operate on systems of first-order equations. To apply them to the algorithm of substitution the higher-order facets must be translated into first-order structures. A partial evaluation system is used to accomplish this.

The software allowing a complete automatic transformation is not yet written. The transformations below have been performed with the Schism partial evaluator [8], the program called Firstify [3] which performs the Reynolds Algorithm [16] and the Astre program transformation system [4].

4.1 Transformation of the *map_with_policy* Operator

The first step is to rewrite the program using the *map_with_policy* operator for the type $Term(\alpha)$ as a system of first-order functions. A partial evaluator can be used to specialize higher-order functions decreasing their order level. For example, consider the particular function σ_0 in the example in Section 1, and the call *apply_substitution* σ_0. A partial evaluator produces a program that does not contain *apply_substitution* in its full generality; it specializes the definition of *apply_substitution* for the particular constant σ_0. This specialization, called

apply_substitution$\_\sigma_0$, does not have a function as an argument, so it is first-order.

Unfortunately, this technique is insufficient for processing calls of *map_with_policy*, which is called twice in the program in Figure 4. The specialization of *map_with_policy* for a particular policy function K and seed function g_0 gives the following function *Mwp_g*:

$$\begin{array}{lcl} Mwp\_g\ (g, Var(n)) & = & Var(g(n)) \\ Mwp\_g\ (g, Abs(t)) & = & Abs(Mwp\_g(K\ g, t)) \\ Mwp\_g\ (g, App(t, t')) & = & App(Mwp\_g(g, t), Mwp\_g(g, t')) \end{array}$$

The function *Mwp_g* has a function as an argument. But if it is specialized for a particular function g_0, the partial evaluator has to specialize the internal call $Mwp\_g(K\ g, t)$; it loops on this attempt. Fortunately, the partial evaluator is able to detect this circumstance, allowing it to select another technique. The alternative technique translates the higher-order functions into a system of first-order functions. This standard encoding, which is due to Reynolds [16], is implemented in a program called Firstify [3]. Let us outline below how it works with the *map_with_policy* operator.

1. The first step constructs a data type that encodes how the higher-order arguments are manipulated and applied. In this case the functions to be encoded are g_0 and $K\ g$. For the constant function, g_0, a constant C is introduced as a summand in the data type *Func*. The argument $K\ g$ cannot be encoded by a simple constant value because it contains g as a free variable. Since g is a higher-order parameter, it will already be represented by a value of type *Func*. Hence the new constructor, F, representing the application of K, must have type $Func \rightarrow Func$. This gives the data type *Func*, defined

   $$\textbf{datatype}\ Func = C \mid F(Func).$$

   The introduction of this type is a rediscovery of the sequence of functions $g_0, g_1, \ldots$ because it encodes each function in the family. The function g_0 is encoded by C, and the function g_3, for example, is encoded by $F(F(F(C)))$, which is written F^3.

2. The functions appearing as actual arguments are replaced by their encodings. The argument functions do not exist anymore—they are replaced by first-order data. In the call $Mwp\_g(g_0, M)$, g_0 is no longer a function but a first-order value, $\lceil g_0 \rceil$, of type *Func*. The definition of Mwp_g leads to the new function Mwp_g':

$$\begin{array}{lcl} Mwp\_g'(\lceil g \rceil, Var(n)) & = & Var(\lceil g \rceil(n)) \\ Mwp\_g'(\lceil g \rceil, Abs(t)) & = & Abs(Mwp\_g'(F(\lceil g \rceil), t)) \\ Mwp\_g'(\lceil g \rceil, App(t, t')) & = & App(Mwp\_g'(\lceil g \rceil, t), Mwp\_g'(\lceil g \rceil, t')) \end{array}$$

   But since $\lceil g \rceil$ is not a function, the application $\lceil g \rceil(n)$ is nonsense.

3. To make sense of the applications of functional parameters in the original programs "application" functions are introduced. Specifically the function $apply\_g$, defined below, decodes applications of the form $\lceil g \rceil(n)$.

$$\begin{array}{rcl} apply\_g(C, n) & = & g_0(n) \\ apply\_g(F(\lceil g \rceil), n) & = & (K\ \lambda n.\, apply\_g(\lceil g \rceil, n))(n). \end{array} \quad (5)$$

   Note that $apply\_g$ is a first-order function because its argument, $\lceil g \rceil$, is an element of the type *Func*. The definition of the policy function K is unfolded to get a first-order expression of $apply\_g(F(\lceil g \rceil), n)$. The definition of Mwp_g' can be completed into:

$$\begin{array}{lcl} Mwp\_g'(\lceil g \rceil, Var(n)) & = & Var(apply\_g(\lceil g \rceil, n)) \\ Mwp\_g'(\lceil g \rceil, Abs(t)) & = & Abs(Mwp\_g'(F(\lceil g \rceil), t)) \\ Mwp\_g'(\lceil g \rceil, App(t, t')) & = & App(Mwp\_g'(\lceil g \rceil\, t), Mwp\_g'(\lceil g \rceil, t')) \end{array}$$

This encoding is done with respect to a specific call of *map_with_policy* $Z\ g_0\ M$. In the program in Figure 4 there are two such calls. The new functions corresponding to Mwp_g and $apply\_g$ constitute a first-order program equivalent to the functions generated by *map_with_policy*.

4.2 Application to *apply_substitution*

Using the preceding techniques, the function *apply_substitution* is successfully transformed into the first-order program in Figure 5. For a given substitution σ_0 , partial evaluation of

```
fun apply_substitution_σ0(M) =
    let  fun apply_f(SUCC, n)       = s(n)
       | apply_f(FSEQ(f), n)        = if n = 0 then 0
                                      else s(apply_f(f, n − 1))
         fun Mwp_f(f, Var(n))       = Var(apply_f(f, n))
       | Mwp_f(f, Abs(t))           = Abs(Mwp_f(FSEQ(f), t))
       | Mwp_f(f, App(t, t′))       = App(Mwp_f(f, t), Mwp_f(f, t′))
         fun apply_σ(S0, n)         = σ0(n)
       | apply_σ(SUBST(σ), n)       = if n = 0 then unit(0)
                                      else Mwp_f(SUCC, (apply_σ(σ, n − 1)))
         fun Mwp_σ(σ, Var(n))       = Var(apply_σ(σ, n))
       | Mwp_σ(σ, Abs(t))           = Abs(Mwp_σ(SUBST(σ), t))
       | Mwp_σ(σ, App(t, t′))       = App(Mwp_σ(σ, t), Mwp_σ(σ, t′))
    in   mult(Mwp_σ(S0,M))
    end
```

Figure 5: First-order Program

an instance *apply_substitution* σ_0 specializes the function *apply_substitution* into a function *apply_substitution_*σ_0. The data type *Subst* and the data type *Fseq* are introduced using the program Firstify which implements Reynolds' techniques for the encodings of *lift* and *shift*.

```
datatype Subst = S0                    datatype Fseq = SUCC
               | SUBST(Subst)                        | FSEQ(Fseq)
```

These two data types are isomorphic to the data type *Nat* which is implemented efficiently in the hardware[2]. However, the specialized function *Mwp_*σ does not exploit the efficient implementation since it uses the (essentially unary) representation of the data type instead. Thus, the function *apply_*σ must peel off all of the data constructors each time *Mwp_*σ is applied to $Var(n)$. For example, after three levels of abstraction, σ_3 is represented by $SUBST(SUBST(SUBST(S0)))$. (The same is also true of the function *Mwp_f*.) To eliminate this inefficiency, which was present in the calling behavior of the original algorithm, the data

[2]The constructors for the data type *Nat* are 0 and s, i.e. **datatype** $Nat = 0 \mid s(Nat)$.

types *Subst* and *Fseq* must be changed to the uniform data type *Nat*. This transformation can be performed automatically by Astre. Ultimately the explicit use of *Nat* will facilitate the use of primitive arithmetic in the program.

5 Simple Transformations

The following two simple transformations are performed automatically by Astre after introducing new function symbols. The first one introduces indexes to count the level of abstractions. The second replaces the composition of *Mwp* with the function *mult* by a single function. The order of these transformations does not matter; they can be done simultaneously.

For technical reasons recursive definitions of the form

$$g(n) = \textbf{if}\, n = 0\, \textbf{then}\, e_1\, \textbf{else}\, e_2$$

are manipulated more effectively by Astre in the equivalent form:

$$\begin{array}{ll} g(0) & = [e_1](n \mapsto 0) \\ g(s(n)) & = [e_2](n \mapsto s(n)) \end{array}$$

This restriction of the form of equations ensures the termination of the rewriting used by Astre to unfold the definition of g.

5.1 Introduction of Indexes

The isomorphism between the automatically generated type *Subst* and the natural numbers is made explicit by introducing the function $iso\_\sigma : Nat \rightarrow Subst$:

$$\begin{array}{lrcl} \textbf{fun} & iso\_\sigma(s(i)) & = & SUBST(iso\_\sigma(i)) \\ \mid & iso\_\sigma(0) & = & S0 \end{array}$$

The functions $apply\_\sigma$ and Mwp_σ are replaced by the new functions $\sigma(i,n)$ (for $\sigma_i(n)$) and Mwp_σ', respectively. These functions satisfy $\sigma(i,n) = apply\_\sigma(iso\_\sigma(i), n)$ and $Mwp\_\sigma'(i,n) = Mwp\_\sigma(iso\_\sigma(i), n)$. Using these new equations, the Astre system implements the data type

```
fun apply_substitution_σ0(M) =
    let  fun f(0, n)                 =  s(n)
           | f(s(i), 0)              =  0
           | f(s(i), s(n))           = s(f(i, n))
         fun Mwp_f'(i, Var(n))       =  Var(f(i, n))
           | Mwp_f'(i, Abs(t))       =  Abs(Mwp_f'(s(i), t))
           | Mwp_f'(i, App(t, t'))   =  App(Mwp_f'(i, t), Mwp_f'(i, t'))
         fun σ(0, n)                 =  σ0(n)
           | σ(s(i), n)              =  unit(0)
           | σ(s(i), s(n))           =  Mwp_f'(0, σ(i, n))
         fun Mwp_σ'(i, Var(n))       =  Var(σ(i, n))
           | Mwp_σ'(i, Abs(t))       =  Abs(Mwp_σ'(s(i), t))
           | Mwp_σ'(i, App(t, t'))   =  App(Mwp_σ'(i, t), Mwp_σ'(i, t'))
    in   mult(Mwp_σ'(0, M))
    end
```

Figure 6: Program with indexes

Subst using the data type *Nat*. New functions to implement the data type *Fseq* using *Nat* are also provided to the Astre system which then gives the program in Figure 6. The program in Figure 6 does not improve the performance of the program in Figure 5. However, its explicit use of numbers is key to the improvements presented in the next section.

5.2 Composition Step

The transformation continues with a simple (automatic) step that replaces the composition of *mult* with Mwp_σ' by a single function.[3] This is accomplished automatically by the introduction of a function symbol, *Ewp*, which is equated to the composition of *mult* with Mwp_σ', i.e., $Ewp(0, M) = mult(Mwp\_\sigma'(0, M))$. Astre gives a program which uses neither *mult*, nor Mwp_σ'

[3] *Ewp* is a mnemonic for extension with policy.

$$
\begin{array}{l}
\textbf{fun } apply\_substitution\_\sigma_0(M) = \\
\quad \textbf{let} \quad \textbf{fun } f(0,n) \qquad\qquad = s(n) \\
\qquad\qquad\quad \mid f(s(i),0) \qquad\qquad = 0 \\
\qquad\qquad\quad \mid f(s(i),s(n)) \qquad = s(f(i,n)) \\
\qquad\qquad \textbf{fun } Mwp(i,\ Var(n)) \quad = \ Var(f(i,n)) \\
\qquad\qquad\quad \mid Mwp(i,Abs(t)) \quad = Abs(Mwp(s(i),t)) \\
\qquad\qquad\quad \mid Mwp(i,App(t,t')) = App(Mwp(i,t),Mwp(i,t')) \\
\qquad\qquad \textbf{fun } \sigma(0,n) \qquad\qquad = \sigma_0(n) \\
\qquad\qquad\quad \mid \sigma(s(i),n) \qquad\qquad = unit(0) \\
\qquad\qquad\quad \mid \sigma(s(i),s(n)) \qquad = Mwp(0,\sigma(i,n)) \\
\qquad\qquad \textbf{fun } Ewp(i,\ Var(n)) \quad = \sigma(i,n) \\
\qquad\qquad\quad \mid Ewp(i,Abs(t)) \quad = Abs(Ewp(s(i),t)) \\
\qquad\qquad\quad \mid Ewp(i,App(t,t')) = App(Ewp(i,t),Ewp(i,t')) \\
\quad \textbf{in} \quad Ewp(0,M) \\
\quad \textbf{end}
\end{array}
$$

Figure 7: Composed Program

that includes the following definition of Ewp:

$$
\begin{array}{l}
\textbf{fun } Ewp(i,\ Var(n)) \quad = \sigma(i,n) \\
\qquad \mid Ewp(i,Abs(t)) \quad = Abs(Ewp(s(i),t)) \\
\qquad \mid Ewp(i,App(t,t')) = App(Ewp(i,t),Ewp(i,t'))
\end{array}
$$

The main body of the function is then replaced by $Ewp(0,M)$. The functions *mult* and Mwp_σ', which have become useless, are removed. Since the Mwp_σ' has now been eliminated, Mwp_f' is renamed Mwp to simplify the nomenclature below.

6 Transformation of the Sequence of the σ Functions

The transformations in this section exploit the arithmetic arguments introduced above to improve the expensive and redundant recursive calculations in σ and Ewp. Indeed, the transformation aims at discovering conditionals and subtraction from a constructor-based definition of a binary arithmetic symbol.

The function $\sigma(i,n)$ of the transformed program is a rediscovery of the series of functions

$\sigma_i(n)$ of Section 1. To further refine this program, a specific instance of *apply_substitution* σ_0 must be specified. In what follows, the substitution function σ_0, needed for the contraction described in Section 1, is used to illustrate the specialization. Recall that σ_0 replaces variables of index 0 with the term $\lambda.0\ 1$, which is represented by $Abs(App(\,Var(0),\,Var(1)))$. Thus, $\sigma_0(0) = Abs(App(\,Var(0),\,Var(1)))$ and $\sigma_0(s(n)) = unit(n)$. Unfolding these equations yields a complete constructor-based definition of $\sigma(i,n)$:

$$\begin{aligned}
&\sigma(0,0) = Abs(App(\,Var(0),\,Var(1)))\\
&\sigma(0,s(n)) = unit(n)\\
&\sigma(s(i),0) = unit(0)\\
&\sigma(s(i),s(n)) = Mwp(0,\sigma(i,n))
\end{aligned} \tag{6}$$

Since the equational program is complete with respect to $Nat * Nat$, the computation of any instance of $\sigma(i,n)$ results in a ground constructor term. For example, $\sigma(4,2)$ yields:

$$\sigma(s(s(s(s(0)))),s(s(0))) \rightarrow \tag{7}$$

$$Mwp(0,\sigma(s(s(s(0))),s(0))) \rightarrow \tag{8}$$

$$Mwp(0,Mwp(0,\sigma(s(s(0)),0))) \rightarrow^* Var(s(s(0)))$$

Rewrites (7) and (8) are unfoldings by equation (6). Computation of any instance of $\sigma(i,n)$ by naturals can begin with unfoldings using (6) until a subterm, $\sigma(u,v)$, in which u *and/or* v are equal to 0 is obtained.

This suggests a target program of the form:

$$\sigma(i,n) = \textbf{if}\ i > n\ \textbf{then}\ e_1\ \textbf{else if}\ i = n\ \textbf{then}\ e_2\ \textbf{else}\ e_3$$

where e_1, e_2, and e_3 are expressions. The transformation will be beneficial if these expressions are efficient. This step introduces a form of function definition by a conditional (instead of structural induction) that violates the technical restriction on programs used to assure termination of rewriting as required by the Astre system. Presently, Astre does not perform this part of the transformation. Moreover, the transformation does not directly generate the conditional; instead it generates the complete definition: $\sigma(s(i)+k,k) = u_1$, $\sigma(k,k) = u_2$ and

$\sigma(k, s(n)+k) = u_3$. This definition, which is no longer constructor-based, is translated directly into a conditional following the pattern above.

6.1 First Transformation Step

The general strategy of the two transformation steps that follow is to discover arithmetic operations implicit in the recursion structure of programs. The goal of the first transformation step is to find the conditional and subtraction from a constructor-based definition of a binary arithmetic symbol which is a simultaneous iterator like σ. Such functions follow the following general pattern for simultaneous iterators:

$$\begin{aligned} &G(0,0) = t \\ &G(s(i),0) = h_1(i) \\ &G(0,s(n)) = h_2(n) \\ &G(s(i),s(n)) = \varphi(G(i,n)) \end{aligned}$$

For example, the constructor-based presentation of the function computing the maximum (or the equality) of two natural numbers follows this general pattern. The first step in this process is a definition that makes the iteration structure of functions explicit. A function G computes $G(6,2)$ as $\varphi(\varphi(G(4,0))) = \varphi^2(h_1(3))$. In the same way, it computes $G(3,7)$ as $\varphi^3(h_2(3))$, and $G(4,4)$ as $\varphi^4(t)$. The results are the same with a function G following the conditional pattern:

$$G(i,n) = \textbf{if}\, i > n\, \textbf{then}\, \varphi^n(h_1(i-n-1))\, \textbf{else if}\, i = n\, \textbf{then}\, \varphi^n(t)\, \textbf{else}\, \varphi^i(h_2(n-i-1))$$

The number k of applications of the function φ denoted by φ^k is made explicit by an index k in the following definition:

Definition 1 *Let x be a variable of type α, let y_i be a term of type β_i for each $i = 1, \cdots, n$, and let φ be a function of type $\beta_1 * \cdots * \alpha * \cdots * \beta_n \to \alpha$. The function $\hat{\varphi}$ of type $Nat * (\beta_1 * \cdots * \alpha * \cdots * \beta_n) \to \alpha$ is defined by:*

$$\begin{aligned} \hat{\varphi}(s(k),(y_1,\cdots,x,\cdots,y_n)) &= \varphi(y_1,\cdots,\hat{\varphi}(k,(y_1,\cdots,x,\cdots,y_n)),\cdots,y_n) \\ \hat{\varphi}(0,(y_1,\cdots,x,\cdots,y_n)) &= x \end{aligned}$$

Proposition 1

$$\hat{\varphi}(k,(y_1,\cdots,\varphi(y_1,\cdots,y,\cdots,y_n),\cdots,y_n)) = \varphi(y_1,\cdots,\hat{\varphi}(k,(y_1,\cdots,y,\cdots,y_n)),\cdots,y_n)$$

Proof: By induction on k. □

An immediate consequence of Definition 1 is $\hat{\varphi}(1,x) = \varphi(x)$, where $x : \beta_1 * \cdots * \alpha * \cdots * \beta_n$.

Having made the iteration structure of functions explicit, the next theorem helps program transformations exploit that structure. To simplify the exposition, consider the case in which $\varphi : \alpha \to \alpha$. In this case $\hat{\varphi} : Nat * \alpha \to \alpha$ and $\hat{\varphi}(k,n) = \varphi^k(x)$, where φ^k denotes k applications of φ. Suppose now that $f : Nat * Nat \to \alpha$ satisfies the equation: $f(s(i),s(n)) = \varphi(f(i,n))$; then $f(4,7) = \varphi^4(f(0,3)) = \hat{\varphi}(4,f(0,3))$. More generally, $f(i+k,n+k) = \hat{\varphi}(k,f(i,n))$. In fact, if $F : Nat * Nat \to \alpha$ then F is a simultaneous iterator if and only if $\hat{\varphi}(k,F(x,y)) = F(x+k,y+k)$, which is the result expressed by Theorem 1.

Theorem 1 *Assume f of type $Nat^n \to \alpha$, let y_i be a term of type β_i for each $i = 1,\cdots,n$, and let φ be a function of type $\beta_1 * \cdots * \alpha * \cdots * \beta_n \to \alpha$. The following are equivalent:*

1. $f(s(x_1),\cdots,s(x_n)) = \varphi(y_1,\cdots,f(x_1,\cdots,x_n),\cdots,y_m)$
2. $\hat{\varphi}(k,(y_1,\cdots,f(x_1,\cdots,x_n),\cdots,y_n)) = f(x_1+k,\cdots,x_n+k)$

Proof: That *1* implies *2* is obvious by instantiating k to 1. The converse is proved by induction on k. □

To apply this theorem to (6), let $Mwp0(x)$ be $Mwp(0,x)$ and introduce the equation:

$$\widehat{Mwp0}(k,\sigma(i,n))) = \sigma(i+k,n+k)$$

This gives the equational definition of $\sigma(i,n)$:

$$\begin{aligned}
&\sigma(s(i)+k,k) = \widehat{Mwp0}(k,\, unit(0)) \\
&\sigma(k,k) = \widehat{Mwp0}(k,\, Abs(App(\,Var(0),\, Var(1)))) \\
&\sigma(k,s(n)+k) = \widehat{Mwp0}(k,\, unit(n))
\end{aligned}$$

This definition can be rewritten in the conditional form described at the beginning of the section with

$$
\begin{aligned}
e_1 &= \widehat{Mwp0}(n, unit(0)) \\
e_2 &= \widehat{Mwp0}(i, Abs(App(\, Var(0),\, Var(1)))) \\
e_3 &= \widehat{Mwp0}(i, unit(n - i - 1))
\end{aligned}
$$

6.2 Second Transformation Step

The second transformation step transforms the expressions e_1, e_2 and e_3. The definition of $\widehat{Mwp0}$ of type $Term \rightarrow Term$, obtained by Definition 1, refers to the (inefficient) function *Mwp0*. To get an efficient program an alternative (but equivalent) definition of $\widehat{Mwp0}$ that does not refer to *Mwp0* must be generated. Theorem 2 addresses this issue.

To introduce Theorem 2, consider the function *upto*. Informally, $upto(i, n) = [i, i+1, \cdots, n]$. The function *upto* satisfies $upto(s(i), s(n)) = map\ s\ upto(i, n)$. Let map_s be the specialization of the definition of *map* by s:

$$
\begin{aligned}
&map\_s\ [\,] = [\,] \\
&map\_s\ (x :: xs) = s(x) :: (map\_s\ xs)
\end{aligned}
$$

The operators [] and :: are the constructors of the data type $List(\alpha)$. By Theorem 1,

$$(\widehat{map\_s})\,(k, upto(i, n)) = (map\_s)^k\ (upto(i, n)) = upto(i + k, n + k)$$

Theorem 2 will yield the following recursive definition of $(map\_s)^k$, (that is of $\widehat{map\_s}$); it does not refer to map_s.

$$
\begin{aligned}
&(map\_s)^k\ [\,] = [\,] \\
&(map\_s)^k\ (x :: xs) = s^k(x) :: ((map\_s)^k\ xs)
\end{aligned}
$$

Note, in this definition $(map\_s)^k$ is the function being defined. It is to be regarded atomically; map_s is neither defined nor referred to.

Theorem 2 *Let y_i be a term of type β_i for each $i = 1, \cdots, n$, let φ be a function of type $\beta_1 * \cdots * \alpha * \cdots * \beta_n \rightarrow \alpha$, and let C be a constructor of type α. The following are equivalent:*

1. $\varphi(y_1, \cdots, C(x_1, \cdots, x_n), \cdots, y_n) = C(\varphi_1(x_1), \cdots, \varphi_n(x_n))$

2. $\hat{\varphi}(k, (y_1, \cdots, C(x_1, \cdots, x_n), \cdots, y_n)) = C(\widehat{\varphi_1}(k, x_1), \cdots, \widehat{\varphi_n}(k, x_n))$

Proof: That *1* implies *2* is obvious by instantiating k to 1. The converse is proved by induction on k. □

If C is a constructor of arity zero, Theorem 2 degenerates to the two equations

$$\varphi(y_1, \cdots, C, \cdots, y_n) = C$$
$$\hat{\varphi}(k, (y_1, \cdots, C, \cdots, y_n)) = C$$

To apply this result to $\widehat{Mwp0}$, recall that $Mwp0(x) = Mwp(0, x)$ and that:

$$\begin{array}{lcl} Mwp(i, Var(n)) & = & Var(f(i, n)) \\ Mwp(i, Abs(t)) & = & Abs(Mwp(s(i), t)) \\ Mwp(i, App(t, t')) & = & App(Mwp(i, t), Mwp(i, t')). \end{array}$$

Introduction of the specializations $f_0(x) = f(0, x)$, and $Mwp1(x) = Mwp(1, x)$ allows the application of Theorem 2, producing:

$$\begin{array}{lcl} \widehat{Mwp0}(k, Var(n)) & = & Var(\widehat{f_0}(k, n)) \\ \widehat{Mwp0}(k, Abs(t)) & = & Abs(\widehat{Mwp1}(k, t)) \\ \widehat{Mwp0}(k, App(s, t)) & = & App(\widehat{Mwp0}(k, s), \widehat{Mwp0}(k, t)). \end{array}$$

It is easy to show that $\widehat{f_0} = \hat{s}$ because $f(0, x) = s(x)$, and that $\hat{s}(k, a) = a + k$ by induction on k. Therefore $\widehat{Mwp0}(k, Var(n)) = Var(\widehat{f_0}(k, n))$, which is equivalent to $Var(\hat{s}(k, n))$, which can be rewritten $Var(n + k)$. Although this appears to have progressed, it is incomplete because $\widehat{Mwp1}$ is still defined in terms of $Mwp1$. Attempts to define $\widehat{Mwp1}$ by this method, however, will require the function $\widehat{Mwp2}$; this would continue forever. Fortunately, there is another way in which Theorem 1 may be applied to (6), yielding the equation $\widehat{Mwp}(k, (0, \sigma(i, n))) = \sigma(i +$

$k, n+k)$. Applying the same transformation as above produces another conditional definition of $\sigma(i,n)$ with $e_1 = unit(n)$, $e_2 = \widehat{Mwp}(i,(0,Abs(App(\,Var(0),\,Var(1)))))$ and $e_3 = unit(n-1)$. Application of Theorem 2 produces a recursive definition of $\widehat{Mwp}$ that does not refer to Mwp:

$$\begin{aligned}
&\widehat{Mwp}(k,(i,\,Var(n))) = \,Var(\hat{f}(k,(i,n))) \\
&\widehat{Mwp}(k,(i,Abs(t))) = Abs(\widehat{Mwp}(k,(s(i),t))) \\
&\widehat{Mwp}(k,(i,App(s,t))) = App(\widehat{Mwp}(k,(i,s)),\widehat{Mwp}(k,(i,t)))
\end{aligned} \tag{9}$$

The transformation is not yet finished. Equation (9) remains to be improved by finding a recursive definition of $\hat{f}$ that does not refer to the function f.

6.3 Transformation of $\hat{f}$

Recall the equations for f:

$$f(0,n) = s(n) \tag{10}$$
$$f(s(i),0) = 0 \tag{11}$$
$$f(s(i),s(n)) = s(f(i,n)) \tag{12}$$

Applying Theorem 2 to (12) yields:

$$\hat{f}(k,(s(i),s(n))) = s(\hat{f}(k,(i,n))). \tag{13}$$

This suggests attempting a conditional definition for $\hat{f}$. Using equations (10), (11), (12), Theorem 2, Theorem 1, and Definition 1 produces:

$$\hat{f}(k,(0,s(n))) = s(\hat{s}(k,n)) = s(n+k) \tag{14}$$
$$\hat{f}(k,(s(i),0)) = 0 \tag{15}$$
$$\hat{f}(k,(0,0)) = k \tag{16}$$

Applying Theorem 1 to (13) gives: $\hat{f}(k,(i+p,n+p)) = \hat{s}(p,\hat{f}(k,(i,n))) = \hat{f}(k,(i,n)) + p$. Applying that to equations (14), (15), (16) produces

$$\begin{aligned}
&\hat{f}(k,(s(i)+p,p)) = p \\
&\hat{f}(k,(p,s(n)+p)) = n+1+k+p \\
&\hat{f}(k,(p,p)) = k+p
\end{aligned}$$

$$
\begin{array}{l}
\textbf{fun } \mathit{apply\_substitution\_}\sigma_0(M) = \\
\quad \textbf{let} \quad \textbf{fun } \widehat{Mwp}(k,(i,\mathit{Var}(n))) \;=\; \textbf{if } i > n \textbf{ then } \mathit{Var}(n) \textbf{ else } \mathit{Var}(n+k) \\
\qquad\quad | \; \widehat{Mwp}(k,(i,\mathit{Abs}(t))) \;=\; \mathit{Abs}(\widehat{Mwp}(k,(s(i),t))) \\
\qquad\quad | \; \widehat{Mwp}(k,(i,\mathit{App}(t,t'))) = \; \mathit{App}(\widehat{Mwp}(k,(i,t)),\widehat{Mwp}(k,(i,t'))) \\
\qquad \textbf{fun } \sigma(i,n) \qquad\qquad\quad \;=\; \textbf{if } i > n \textbf{ then } \mathit{unit}(n) \\
\qquad\qquad\qquad\qquad\qquad\quad\; \textbf{else if } i = n \textbf{ then} \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad \widehat{Mwp}(i,(0,\mathit{Abs}(\mathit{App}(\mathit{Var}(0),\mathit{Var}(1))))) \\
\qquad\qquad\qquad\qquad\qquad\quad\; \textbf{else } \mathit{unit}(n-1) \\
\qquad \textbf{fun } \mathit{Ewp}(i,\mathit{Var}(n)) \qquad\;\; \;=\; \sigma(i,n) \\
\qquad\quad | \; \mathit{Ewp}(i,\mathit{Abs}(t)) \qquad\;\; \;=\; \mathit{Abs}(\mathit{Ewp}(s(i),t)) \\
\qquad\quad | \; \mathit{Ewp}(i,\mathit{App}(t,t')) \qquad \;=\; \mathit{App}(\mathit{Ewp}(i,t),\mathit{Ewp}(i,t')) \\
\quad \textbf{in} \quad \mathit{Ewp}(0,M) \\
\quad \textbf{end}
\end{array}
$$

Figure 8: Final result

This equational definition is equivalent to the program:

$$\hat{f}(k,(i,n)) = \textbf{if } i > n \textbf{ then } n \textbf{ else if } i = n \textbf{ then } n+k \textbf{ else } n+k.$$

The program simplifies to: $\hat{f}(k,(i,n)) = \textbf{if } i > n \textbf{ then } n \textbf{ else } n+k$. By unfolding $\hat{f}$ and by a well known property of the conditional, equation (9) becomes:

$$\widehat{Mwp}(k,(i,\mathit{Var}(n))) = \textbf{if } i > n \textbf{ then } \mathit{Var}(n) \textbf{ else } \mathit{Var}(n+k)$$

Including the transformed form of σ, which comes from above, produces the program in Figure 8 which does not perform redundant computations for σ_i and f_i. The transformation involved in this section has been done manually. However the transformation process is systematic and involves equational reasoning using Theorem 1 and Theorem 2. It shows implicitly how to automatically transform a constructor-based definition of a simultaneous iterator function of type $\mathit{Nat} * \mathit{Nat} \rightarrow \mathit{Nat}$ into a more efficient conditional form.

7 Reuse

Although this paper has focused on the λ-calculus, the specification can be applied to virtually any abstract syntax with a regular binding structure, provided its type can be expressed as a monad and the appropriate definition of *map_with_policy* can be given. Figure 9 illustrates this by showing how the function defined in Figure 4 can be expressed in a Standard ML functor abstracted on the signature of a monadic abstract syntax. Given this abstract presentation, the substitution algorithm can be specialized to a new abstract syntax simply by Standard ML functor application. For example, the structure encoding the enrichment of λ-terms with *let* is given in Figure 10. In this case, *map_with_policy* must apply Z to f when it enters the component in which the bound variable has been introduced. This ability to reuse specifications is one of the strongest arguments for the adoption of monads as a tool to structure program development.

The transformation of the apply_substitution function on the enriched abstract syntax is essentially the same as that presented above. A simple replay mechanism should be sufficient to perform the transformation of the enriched program.

8 Directions

The program development in this paper illustrates several new techniques. It makes the monadic structure in the development of the algorithm explicit. It supports this structure with new program transformation techniques that allow the implicit use of arithmetic to be "rediscovered" formally. It demonstrates the feasibility of integrating tools for monadic program development, which tend to be higher-order, with relatively standard program transformation technology, which is strictly first-order.

```
signature TermMonad =
  sig
    type 'a Term
    val unit : 'a -> 'a Term
    val mult : 'a Term Term -> 'a Term
    val map : ('a -> 'b) -> 'a Term -> 'b Term
    val map_with_policy : (('a -> 'b) -> 'a -> 'b)
                          -> ('a -> 'b) -> 'a Term -> 'b Term
    val extension_with_policy : (('a -> 'b Term)
                                  -> 'a -> 'b Term)
                                -> ('a -> 'b Term)
                                   -> 'a Term -> 'b Term
  end


functor SubstitutionFunctor(T: TermMonad) =
    struct
        open T
        fun apply_substitution sigma_0 M
            = let fun succ x = x+1
                  fun lift f
                      = fn n => if n = 0 then 0 else 1+f(n-1)
                  fun shift sigma
                      = fn n => if n = 0 then unit 0
                                else map_with_policy lift succ (sigma (n-1))
              in extension_with_policy shift sigma_0 M
              end;
    end
```

Figure 9: Standard ML functor for an arbitrary abstract syntax.

```
structure Let:TermMonad =
    struct
        datatype 'a Term =
            Var of 'a
          | App of 'a Term * 'a Term
          | Abs of 'a Term
          | Let of 'a Term * 'a Term;

        val unit = Var

        fun mult (Var t) = t
          | mult (App(a,b)) = App(mult a, mult b)
          | mult (Abs(a)) = Abs(mult a)
          | mult (Let(a,b)) = Let(mult a,mult b);

        fun map f (Var a) = Var (f a)
          | map f (App(a,b)) = App(map f a,map f b)
          | map f (Abs(a)) = Abs(map f a)
          | map f (Let(a,b)) = Let(map f a,map f b);

        fun map_with_policy Z f (Var a) = Var (f a)
          | map_with_policy Z f (App(a,b))
            = App(map_with_policy Z f a,map_with_policy Z f b)
          | map_with_policy Z f (Abs(a))
            = Abs(map_with_policy Z (Z f) a)
          | map_with_policy Z f (Let(a,b))
            = Let(map_with_policy Z f a,map_with_policy Z (Z f) b);

        fun extension_with_policy Z f = mult o (map_with_policy Z f);
    end
```

Figure 10: Structure encoding the abstract syntax enriched with *let*

We believe this is a general paradigm for automated program development. We are developing a transformation system for functional programs that is a combination of automated strategies implemented by different tools[2]. The transformation is performed according to the following scheme: (1) conversion of the program into a first-order set of constructor-based equations E (see Section 4), (2) manipulation of E by rewriting techniques to automate different strategies based on the unfold/fold method [7] (see Section 5), (3) translation of the constructor-based set of equations to an efficient functional program, e.g. introduction of conditionals and machine arithmetic. The case study presented here is a good (though not typical) example of this process.

The system we are developing includes an algebraic programming notation with limited support for monads[13], transformation tools for conversion to first-order (including specialization and Reynolds' defunctionalization[3]), and automatic first-order transformation by rewriting techniques[4].

However, some important pieces that are required for automating the formal methods presented in this paper are not yet implemented. For example, support of the rich monadic development presented in Section 1 and the use of arithmetic introduced by the transformations presented in Section 6 exceed the current functionality of our system.

The paper has presented a clearly motivated and correct program development for a substitution algorithm for λ-terms. It has taken an abstract algorithm, with extensive use of higher-order concepts, reduced it to a first-order program, introduced index arithmetic and produced an efficient algorithm that exploits computer arithmetic.

We would like to thank Richard Kieburtz, Jeffrey Bell, and our other colleagues in the Pacific Software Research Center at the Oregon Graduate Institute of Science & Technology. We also wish to thank the referees for their constructive and insightful remarks.

A Correctness of the substitution algorithm

This section demonstrates the equivalence of the substitution function defined in Figure 4 to the definition of substitution in the calculus of explicit substitutions of Curien, Hardin and Lévy [10, 9]. In Curien, Hardin, and Lévy's calculus with explicit substitutions, as well as in the other calculi proposed by Abadi, Cardelli, Curien and Lévy [1] and Lescanne [14], the result of applying the substitution σ_0 to the term M is translated by a *closure*: $M[mult\ (map\ \sigma_0)]$. Note that these calculi represents substitution as functions from terms to terms, hence we use the natural extension of σ_0 rather than σ_0 itself. To prove the correctness of the substitution algorithm, it suffices to prove that $apply\_substitution\ \sigma_0\ M$ computes $M[mult\ (map\ \sigma_0)]$.

In the proof, three equivalence relations will be used: the symbol $=$ will be used for the equivalence of ML expressions, the symbol $\leftrightarrow$ will be used for equivalence in the calculus of explicit substitutions, and the symbol $\sim$ will be used for a semantic equivalence relating them. This semantic equivalence is generated by identity of results of computations and by substitution of equals for equals.

Thus, we must show that: $mult\ (map\_with\_policy\ shift\ \ \sigma_0\ M) \sim M[mult\ (map(\sigma_0)]$, that is, by definition of map_with_policy, prove:

$$mult\ (map(\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M) \sim\ \ M[mult\ (map(\sigma_0)$$

The definition of a closure, adapted from Curien, Hardin and Lévy, is:

$$\begin{array}{l} App(M,N)[s] \ \leftrightarrow\ App(M[s], N[s]) \\ (Abs\ M)[s] \ \leftrightarrow\ (Abs\ M)[\Uparrow s] \\ (Var\ n)[s] \ \leftrightarrow\ (Var\ n) \end{array}$$

where $s = mult\ (map\ \sigma_0)$ and, $\Uparrow$ is defined as:

$$\begin{array}{l} \Uparrow s(Var\ 0) \ \leftrightarrow\ (Var0) \\ \Uparrow s(Var\ (n+1)) \ \leftrightarrow\ (s(Var\ n))[\uparrow] \end{array}$$

with $\uparrow = map\ succ$.

The proof uses the following **shift lemma**:

$$shift\ \ \sigma_0 \sim \Uparrow\ (mult\ (map\ \sigma_0)).$$

whose proof uses a second lemma, the **lift lemma**:

$$M[map\ f_0] \sim\ \ map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ M.$$

Proof of the **lift lemma**:

Proof: The proof is by structural induction on the structure of terms. There are three cases:

1. The term M is a variable $(Var\ n)$:

   $(Var\ n)[map\ f_0] \leftrightarrow\ (map\ f_0)(Var\ n)$ by definition of the closure
   $=\ Var\ (f_0\ n)$ by definition of *map*
   $=\ map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ (Var\ n)$ by definition of *map*.

2. The term M is an application $App(M, N)$:

   $map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots]\ App(M, N)$
   $=\ App(map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ M, map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ N$
   by definition of *map*
   $\sim\ App(M[map\ f_0], N[map\ f_0])$ by induction
   $\leftrightarrow\ App(M, N)[map\ f_0]$ by definition of the closure.

3. The term M is an abstraction $Abs(M)$: On one hand,

   $map(f_0, lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ Abs(M)$
   $=\ Abs(map(lift\ \ f_0, lift^{\,2}\ f_0, \ldots)\ M)$ by definition of *map*
   $\sim\ Abs(M[map\ (lift\ \ f_0)])$ by induction.

   On another hand,

   $Abs(M)[map\ f_0] \leftrightarrow\ Abs(M[\Uparrow (map\ f_0)])$ by definition of the closure.

The proof is done if

$$\Uparrow\ (map\ f_0)\ \sim\ map\ (lift\ \ f_0).$$

which can be easily proved by mathematic induction on the de Bruijn number n of a variable using the definition of $\Uparrow$ and *lift*. □

Proof of the **shift lemma**:

Proof: The proof is by mathematical induction on the de Bruijn number n of a variable.

1. The de Bruijn number is 0:

$$shift\ \ \sigma_0 0\ =\ (Var\ 0) \text{ by definition of } shift,$$

and

$$\Uparrow\ (mult\ (map\ \sigma_0))(Var\ 0)\ =\ (Var\ 0) \text{ by definition of } \Uparrow.$$

2. The de Bruijn number is $n+1$: On one hand,

$$\begin{array}{l} shift\ \ \sigma_0\ (n+1)\ =\ map\ (succ, lift\ \ succ, lift^{\,2}\ succ, \ldots)\ (\sigma_0\ n) \\ \text{by definition of } shift \\ \sim\ (\sigma_0 n)[map\ succ] \text{ by the } \textbf{lift lemma}. \end{array}$$

On another hand,

$$\begin{array}{l} \Uparrow\ (mult\ (map\ \sigma_0))(Var(n+1)) \leftrightarrow mult\ (map\ \sigma_0(Var\ n))[\uparrow] \\ \text{by definition of } \Uparrow \\ =\ mult(Var\ (\sigma_0\ n))[\uparrow] \text{ by definition of } map \\ =\ (\sigma_0\ n)[\uparrow] \text{ by definition of } mult. \end{array}$$

□

Finally, the proof of correctness of the substitution function:

$$mult\ (map(\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M) \sim\ \ M[mult\ (map(\sigma_0)].$$

Proof: The proof is by structural induction on the term structure. There are three cases:

1. The term M is a variable $(Var\ n)$: On one hand,

$$\begin{array}{l} (Var\ n)[mult\ (map\ \sigma_0)] \\ \leftrightarrow\ mult\ (map\ (\sigma_0\ (Var\ n))) \text{ by definition of the closure} \\ =\ mult\ (\sigma_0\ n) \text{ by definition of } map\ =\ \sigma_\circ\ n \text{ by definition of } mult. \end{array}$$

On the other hand,

$$\begin{array}{l} mult\ (map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ (Var\ n)) \\ =\ mult\ (Var(\sigma_0\ n)) \text{ by definition of } map \\ =\ \sigma_0\ n \text{ by definition of } mult \end{array}$$

2. The term M is an application $App(M, N)$: On one hand,

$App(M, N)[mult\ (map\ \sigma_0)]$
$\leftrightarrow\ App(M[mult\ (map\ \sigma_0)], N[mult\ (map\ \sigma_0)])$ by definition of the closure.

On the other hand,

$mult\ (map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ (App(M, N)))$
$=\ mult\ App(map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M$
$\qquad , map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ N)$
by definition of map
$=\ App(mult\ (map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M),$
$\quad mult\ (map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ N))$ by definition of $mult$
$\sim\ App(M[mult\ (map\ \sigma_0)], N[mult\ (map\ \sigma_0)])$ by induction.

3. The term M is an abstraction $Abs\ M$: On one hand,

$Abs(M)[mult\ (map\ \sigma_0)]$
$\sim\ Abs(M[\Uparrow\ (mult\ (map\ \sigma_0))]$ by definition of the closure.

On another hand,

$mult\ (map\ (\sigma_0, shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ (Abs\ M))$
$=\ mult\ (Abs\ (map\ (shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M))$ by definition of map
$=\ Abs\ (mult\ (map\ (shift\ \ \sigma_0, shift^{\,2}\ \sigma_0 \ldots)\ M))$ by definition of $mult$
$\sim\ Abs(M[shift\ \ \sigma_0])$ by induction
$\leftrightarrow\ Abs(M[\Uparrow\ (mult\ (map\ \sigma_0))]$ by the shift lemma.

□

References

[1] M. Abadi, L. Cardelli, P. L. Curien, and J. J. Lévy. Explicit substitutions. *J. of Functional Programming*, 1:375–416, 1991.

[2] Jeffrey Bell et al. Software design for reliability and reuse: A proof-of-concept demonstration. Technical report, Department of Computer Science and Engineering, Oregon Graduate Institute, March 1994.

[3] Jeffrey M. Bell. An implementation of Reynold's defunctionalization method for a modern functional language. Master's thesis, Oregon Graduate Institute, January 1994.

[4] Françoise Bellegarde. Program transformation and rewriting. In *Proceedings of the fourth conference on Rewriting Techniques and Applications*, volume 488 of *LNCS*, pages 226–239, Berlin, 1991. Springer-Verlag.

[5] N. G. de Bruijn. Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem. *Indagaciones Mathematische*, 34:381–392, 1972. Also appeared in the Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam, series A, 75(5).

[6] N. G. de Bruijn. Lambda calculus with namefree formulas involving symbols that represent reference transforming mappings. In *Proceedings of the Koninklijke Nederlandse Akaemie van Wetenschappen*, pages 348–356, Amsterdam, series A, volume 81(3), September 1978.

[7] R. M. Burstall and J. Darlington. A transformation system for developing recursive programs. *Journal of the ACM*, 24:44–67, 1977.

[8] Charles Consel and Olivier Danvy. Tutorial notes on partial evaluation. In *Conference Record of the Twentieth Annual ACM Symposium on Principles of Programming Languages*, pages 493–501. ACM, 1993.

[9] P. L. Curien, Th. Hardin, and J. J. Lévy. Confluence properties of weak and strong calculi of explicit substitutions. Technical Report RR 1617, INRIA, Rocquencourt, 1992.

[10] Th. Hardin and J.J. Lévy. A confluent calculus of substitutions. In *France-Japan Artificial Intelligence and Computer Science Symposium*, 1989.

[11] James Hook, Richard Kieburtz, and Tim Sheard. Generating programs by reflection. Technical Report 92-015, Department of Computer Science and Engineering, Oregon Graduate Institute, July 1992.

[12] Richard B. Kieburtz. A generic specification of prettyprinters. Technical Report CSE-91-020, Department of Computer Science and Engineering, Oregon Graduate Institute, 1991.

[13] Richard B. Kieburtz and Jeffrey Lewis. Algebraic design language (preliminary definition). Technical report, Department of Computer Science and Engineering, Oregon Graduate Institute, January 1994.

[14] P. Lescanne. From $\lambda\sigma$ to λv, a journey through calculi of explicit substitutions. In *Symposium on Principles of Programming Languages*, 1994.

[15] Eugenio Moggi. Notions of computations and monads. *Information and Computation*, 93(1):55–92, July 1991.

[16] John C. Reynolds. Definitional interpreters for higher-order programming languages. In *ACM National Conference*, pages 717–740. ACM, 1972.

[17] Philip Wadler. Theorems for free! In *Proc. of 4th ACM Conf. on Functional Programming Languages and Computer Architecture*, pages 347–359. ACM Press, September 1989.

[18] Philip Wadler. Comprehending monads. In *Proc. 1990 ACM Conference on Lisp and Functional Programming*, pages 61–78, 1990.

[19] Philip Wadler. The essence of functional programming. In *Conference Record of the Nineteenth Annual ACM Symposium on Principles of Programming Languages*. ACM Press, January 1992.